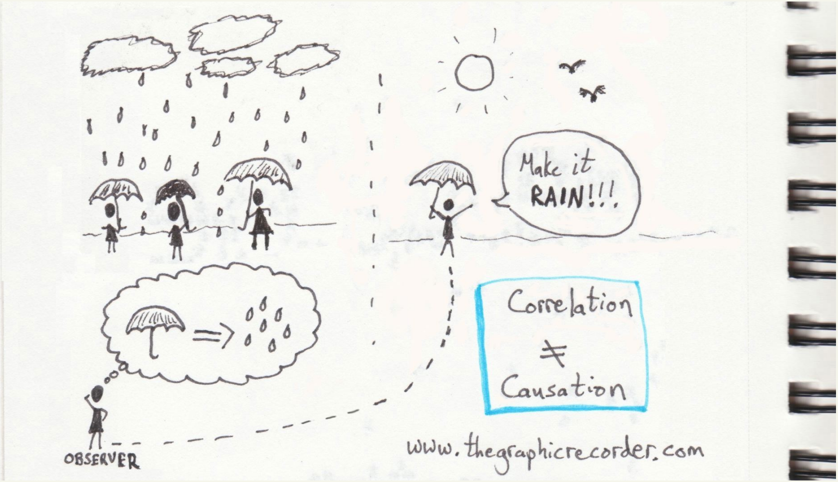
Modelos Bayesianos

Introducción a la Inteligencia Artificial

Agenda

Hablemos de causalidad



Causalidad: ¿Quién causa a quién?

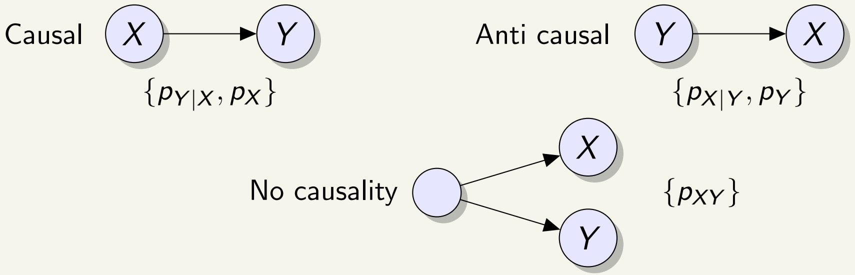
Independent Causal Mechanisms (ICM) Principle

The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

Causalidad: ¿Quién causa a quién?

Independent Causal Mechanisms (ICM) Principle

The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.



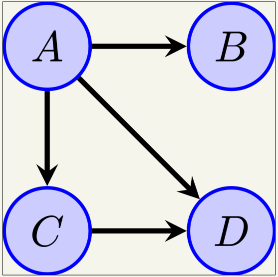
Redes Bayesianas

Modelos Gráficos

Modelos probabilísticos capaz de representarse con un grafo.

Red Bayesiana

Grafo acíclico dirigido que representa la relación de causalidad e independencia de sus variables. Dos variables aleatorias cualesquiera son condicionalmente independientes dados los valores de sus padres causales (y por lo tanto las raices son independientes).



$$p(A, B, C, D) = p(A) \cdot p(B|A) \cdot p(C|A) \cdot p(D|A, C)$$

Inferencia Bayesiana

- Los parámetros θ deben ser considerado realizaciones de una variable aleatoria T con una distribución a priori conocida $p(\theta)$.
- Las muestras son i.i.d. **cuando** se conoce el parámetro.
- La distribución a posteriori de los parámetros se calcula como:

$$p(\theta|\mathcal{D}_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

  con $\mathcal{D}_n = \{x_1, \cdots, x_n\}$.
- Como estimador puntual suele elegirse el *maximo a posteriori* (Θ discreto) y el estimador bayesiano o *media a posteriori* (Θ continuo).
- No son necesarios los estimadores puntuales para predecir:

$$p(x_{\text{test}}|\mathcal{D}_n) = \int_{\Theta} p(x_{\text{test}}|\theta) p(\theta|\mathcal{D}_n) d\theta = \mathbb{E}[p(x_{\text{test}}|T)|\mathcal{D}_n]$$

Inferencia Bayesiana

Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

Inferencia Bayesiana

Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

No confunduir Bayesiano con Relativista

¿Es posible entonces alcanzar verdades en las ciencias empíricas en las que es inevitable decir "no sé"? Sí. Podemos evitar mentir: maximizando incertidumbre (no afirmar más de lo que se sabe) dada la información disponible (sin ocultar lo que sí se sabe).

Inferencia Bayesiana

Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

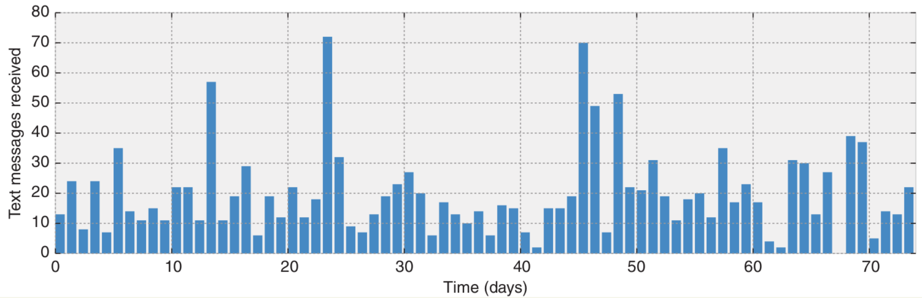
No confunduir Bayesiano con Relativista

¿Es posible entonces alcanzar verdades en las ciencias empíricas en las que es inevitable decir "no sé"? Sí. Podemos evitar mentir: maximizando incertidumbre (no afirmar más de lo que se sabe) dada la información disponible (sin ocultar lo que sí se sabe).

¿Son prácticos los métodos Bayesianos?

Si, no solo por poder *adaptarse* a intentar resolver los mismos problemas que la estadística frecuentista (por ejemplo predicciones), sino que también pueden intentar resolver problemas donde la estadística clásica es insuficiente o iluminar el sistema subyacente con un modelado más flexible.

Ejemplo de modelado Bayesiano



Problema

Un usuario proporciona una serie de recuentos diarios de mensajes de whatsapp enviados. Tiene curiosidad por saber si los hábitos de envío de mensajes han cambiado con el tiempo. ¿Cómo puedes modelar esto?

Ejemplo de modelado Bayesiano

- La cantidad de mensajes en un día deberá ser modelada como una variable discreta cuyos átomos es $\mathbb{N}_0$. Por ejemplo $X_i \sim \text{Poi}(\lambda_i)$.
- Si observamos los datos, parecería que el valor de λ_i aumenta en algún momento durante las observaciones. ¿Cómo podemos representar matemáticamente esta observación? Supongamos que algún día τ durante el período de observación, el parámetro λ_i se incrementa repentinamente. Entonces realmente tenemos dos tasas: una para el período anterior a τ y otro para el resto del período:

$$\lambda_i = \begin{cases} \beta_1 & i < \tau \\ \beta_2 & i \geq \tau \end{cases}$$

- Tanto β_1 como β_2 toman valores reales no negativos. Por ejemplo $\beta_1, \beta_2 \sim \mathcal{E}(\alpha)$. Nuestra estimación de α no influye demasiado en el modelo, por lo que tenemos cierta flexibilidad en nuestra elección. Para evitar ser demasiado obstinados con este parámetro se sugiere:

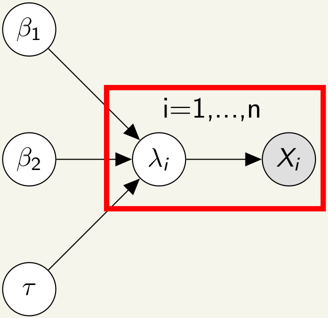
$$\alpha \approx \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

Ejemplo de modelado Bayesiano

- ¿Qué pasa con τ? Debido a la varianza de los datos, es difícil caracterizarlo en detalle. En cambio, podemos asignar la creencia menos informativa posibles $\tau \sim \mathcal{U}\{1 : T\}$.

Ejemplo de modelado Bayesiano

- ¿Qué pasa con τ? Debido a la varianza de los datos, es difícil caracterizarlo en detalle. En cambio, podemos asignar la creencia menos informativa posibles $\tau \sim \mathcal{U}\{1 : T\}$.



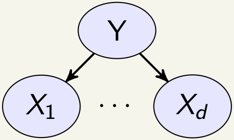
Outline

Naive Bayes

Naive Bayes

Estimar los parámetros (máxima verosimilitud o bayesiano) asumiendo que una relación de causalidad $Y \to X$ con las diferentes componentes $X_j | Y = k$ independientes.

Red Bayesiana



Cálculo

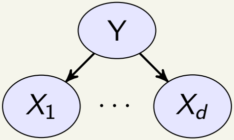
$$p(y|\mathrm{x}) = \frac{p(y) \prod_{j=1}^{d} p(x_j|y)}{p(\mathrm{x})}$$

Naive Bayes

Naive Bayes

Estimar los parámetros (máxima verosimilitud o bayesiano) asumiendo que una relación de causalidad $Y \to X$ con las diferentes componentes $X_j|Y = k$ independientes.

Red Bayesiana



Cálculo

$$p(y|\mathrm{x}) = \frac{p(y)\prod_{j=1}^{d} p(x_j|y)}{p(\mathrm{x})}$$

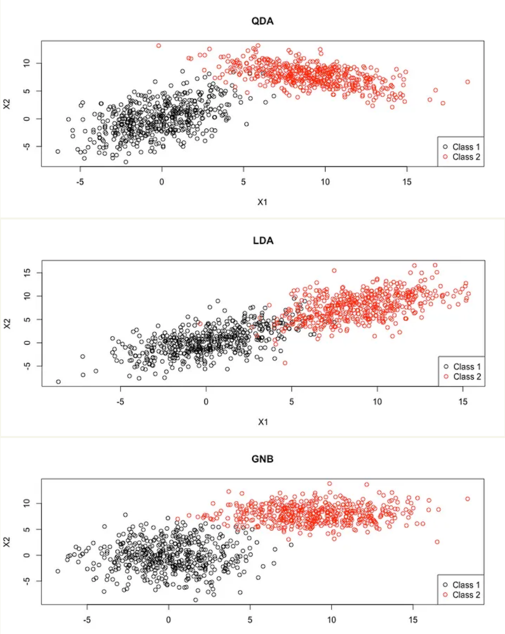
Gaussian Naive Bayes

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{(\mu_1^{(k)}, \sigma_1^{2(k)}), \cdots (\mu_d^{(k)}, \sigma_d^{2(k)})\}$, se modela $Y \sim \mathrm{Cat}(\{c_1, \cdots, c_K\})$ y $X_j|Y = k \sim \mathcal{N}(\mu_j^{(k)}, \sigma_j^{2(k)})$, para luego estimar los parámetros.

Gaussian Naive Bayes (GNB)

Diferencias entre QDA, LDA y GNB

- QDA acepta como Σ_k cualquer conjunto de matrices definidas positiva.
- LDA acepta como Σ_k cualquier matriz pero todas iguales.
- GNB permite tener matrices Σ_k diferentes pero todas diagonales.



Gaussian Naive Bayes (GNB)

QDA

$$\Sigma_k = \frac{1}{|\mathcal{D}_k| - 1} \sum_{x \in \mathcal{D}_k} (x - \mu^{(k)}) \left(x - \mu^{(k)}\right)^T$$

LDA

$$\Sigma = \frac{1}{n - K} \sum_{k=1}^{K} (|\mathcal{D}_k| - 1) \Sigma_k$$

GNB

$$\Sigma_k = \mathrm{DIAG}\left(\sigma_1^{2(k)}, \cdots, \sigma_d^{2(k)}\right), \qquad \sigma_j^{2(k)} = \frac{1}{|\mathcal{D}_k| - 1} \sum_{x \in \mathcal{D}_k} (x_j - \mu_j^{(k)})^2$$

Outline

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB)

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\}$, se modela como $Y \sim \mathrm{Cat}(\{c_1, \cdots, c_K\})$ y $X_j | Y = k \sim \mathrm{Cat}(\{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\})$, utilizando estimadores puntuales.

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB)

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\}$, se modela como $Y \sim \text{Cat}(\{c_1, \cdots, c_K\})$ y $X_j|Y = k \sim \text{Cat}(\{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\})$, utilizando estimadores puntuales.

Inferencia

$$p(y|\mathrm{x}) \propto c_y \prod_{j=1}^{d} \theta_{x_j}^{(y)} = c_y \prod_{m=1}^{V} \left(\theta_m^{(y)}\right)^{N_m}$$

con N_m : Cantidad de predictores con valor m.

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB)

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\}$, se modela como $Y \sim \mathrm{Cat}(\{c_1, \cdots, c_K\})$ y $X_j | Y = k \sim \mathrm{Cat}(\{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\})$, utilizando estimadores puntuales.

Inferencia

$$p(y|\mathrm{x}) \propto c_y \prod_{j=1}^{d} \theta_{x_j}^{(y)} = c_y \prod_{m=1}^{V} \left(\theta_m^{(y)}\right)^{N_m}$$

con N_m : Cantidad de predictores con valor m.

Sobre las variables contadoras

Sea $\mathrm{N} = (N_1, \cdots, N_V)$, es sencillo notar que $\sum_{m=1}^{V} N_m = d$ y $\mathrm{N}|_{Y=k} \sim \mathcal{M}_n(d, [\theta_1^{(k)}, \cdots, \theta_V^{(k)}])$.

Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg\max_{y} \; p(y|\mathrm{x}) = \arg\max_{y} \; \log(c_y) + \sum_{m=1}^{V} N_m \log\left(\theta_m^{(y)}\right)$$

Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg\max_y \; p(y|\mathrm{x}) = \arg\max_y \; \log(c_y) + \sum_{m=1}^{V} N_m \log\left(\theta_m^{(y)}\right)$$

Probabilidades de las Clases

Los parámetros $c_1, \cdots, c_K$ son estimados por máxima verosimilitud como:

$$\hat{c}_k = \frac{\#\{y_i = k\}}{n}$$

Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg\max_y \; p(y|\mathrm{x}) = \arg\max_y \; \log(c_y) + \sum_{m=1}^{V} N_m \log\left(\theta_m^{(y)}\right)$$

Probabilidades de las Clases

Los parámetros $c_1, \cdots, c_K$ son estimados por máxima verosimilitud como:

$$\hat{c}_k = \frac{\#\{y_i = k\}}{n}$$

Sobre el valor d

Cada muestra puede poseer un valor de d diferente. Eso es típico en texto, donde cada documento posee una cantidad diferente de palabras.

Multinomial Naive Bayes

Estimación de $\theta_m^{(k)}$

Se cuenta con datos $\{(\mathrm{N}_i, y_i)\}_{i=1}^n$. Sin embargo, para cada clase k se utilizarán solamente los datos con $\{y_i = k\}$ distribuidos como una multinomial de probabilidades $\theta_1^{(k)}, \cdots, \theta_V^{(k)}$. A su vez, dado que las variables N_m cuentan ocurrencias, puedo compactar todas las muestras de entrenamiento de cada clase en una sola (suficiencia estadística).

$$\tilde{N}_m^{(k)} = \sum_{i=1}^{n} N_{i,m} \cdot \mathbb{1}\{y_i = k\}$$

Multinomial Naive Bayes

Estimación de $\theta_m^{(k)}$

Se cuenta con datos $\{(\mathrm{N}_i, y_i)\}_{i=1}^n$. Sin embargo, para cada clase k se utilizarán solamente los datos con $\{y_i = k\}$ distribuidos como una multinomial de probabilidades $\theta_1^{(k)}, \cdots, \theta_V^{(k)}$. A su vez, dado que las variables N_m cuentan ocurrencias, puedo compactar todas las muestras de entrenamiento de cada clase en una sola (suficiencia estadística).

$$\tilde{N}_m^{(k)} = \sum_{i=1}^n N_{i,m} \cdot \mathbb{1}\{y_i = k\}$$

Modelado: Estimador Bayesiano

Como modelado para el entrenamiento se supone $\mathrm{T} \sim \mathrm{Dir}([\alpha_1, \cdots, \alpha_V])$ y $(\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)})|_{\mathrm{T}=\Theta} \sim \mathcal{M}_n(\tilde{d}^{(k)}, [\theta_1^{(k)}, \cdots, \theta_V^{(k)}])$.

Multinomial Naive Bayes

Dirichlett Distribution

El vector aleatorio $(T_1, \cdots, T_V) \sim \mathrm{Dir}([\alpha_1, \cdots, \alpha_V])$ puede ser pensado como una beta multivariada. Su densidad es de la forma

$$p(\theta_1, \cdots, \theta_V) = \frac{\prod_{m=1}^{V} \Gamma(\alpha_m)}{\Gamma\left(\sum_{m=1}^{V} \alpha_m\right)} \left(\prod_{m=1}^{V} \theta_m^{\alpha_m - 1} \right) \cdot \mathbb{1} \left\{ \sum_{m=1}^{V} \theta_m = 1, \theta_m \geq 0 \right\}$$

con sus marginales $T_m \sim \beta(\alpha_m, \sum_{\eta \neq m} \alpha_\eta)$.

Sobre la beta

Recordar que si $T \sim \beta(a, b)$, entonces $\mathbb{E}[T] = \frac{a}{a+b}$.

Multinomial Naive Bayes

Distribución a Posteriori

$$
\begin{aligned}
& p\left(\theta_1^{(k)}, \cdots, \theta_V^{(k)} | \tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}\right) \\
& \quad \propto P\left(\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)} | \theta_1^{(k)}, \cdots, \theta_V^{(k)}\right) \cdot p\left(\theta_1^{(k)}, \cdots, \theta_V^{(k)}\right) \\
& \quad \propto \left(\prod_{m=1}^{V} (\theta_m^{(k)})^{\tilde{N}_m^{(k)}}\right) \left(\prod_{m=1}^{V} (\theta_m^{(k)})^{\alpha_m - 1} \cdot \mathbb{1}\left\{\theta_m^{(k)} \geq 0\right\}\right) \cdot \mathbb{1}\left\{\sum_{m=1}^{V} \theta_m^{(k)} = 1\right\}
\end{aligned}
$$

con lo cuál $\mathrm{T}|_{\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}} \sim \mathrm{Dir}([\tilde{N}_1^{(k)} + \alpha_1, \cdots, \tilde{N}_V^{(k)} + \alpha_V])$

Multinomial Naive Bayes

Distribución a Posteriori

$$p\left(\theta_1^{(k)}, \cdots, \theta_V^{(k)} | \tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}\right)$$
$$\propto P\left(\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)} | \theta_1^{(k)}, \cdots, \theta_V^{(k)}\right) \cdot p\left(\theta_1^{(k)}, \cdots, \theta_V^{(k)}\right)$$
$$\propto \left(\prod_{m=1}^{V} (\theta_m^{(k)})^{\tilde{N}_m^{(k)}}\right) \left(\prod_{m=1}^{V} (\theta_m^{(k)})^{\alpha_m - 1} \cdot \mathbb{1}\left\{\theta_m^{(k)} \geq 0\right\}\right) \cdot \mathbb{1}\left\{\sum_{m=1}^{V} \theta_m^{(k)} = 1\right\}$$

con lo cuál $\mathrm{T}|_{\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}} \sim \mathrm{Dir}([\tilde{N}_1^{(k)} + \alpha_1, \cdots, \tilde{N}_V^{(k)} + \alpha_V])$

Estimador Bayesiano

$$\hat{\theta}_m^{(k)} = \mathbb{E}[T_m | \tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}] = \frac{\tilde{N}_m^{(k)} + \alpha_m}{\sum_{\eta=1}^{V} \tilde{N}_\eta^{(k)} + \alpha_\eta}$$

Outline

Monte-Carlo

Ley de los grandes números

Sean X_i variables aleatorias i.i.d. con esperanza finita $\mathbb{E}[X]$, entonces $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X]$ (w.p.1).

Monte-Carlo

Ley de los grandes números

Sean X_i variables aleatorias i.i.d. con esperanza finita $\mathbb{E}[X]$, entonces $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X]$ (w.p.1).

Método de Monte-Carlo

Sean X_i variables aleatorias i.i.d. con pdf $p(x)$ o pmf $P(x)$, entonces

$$\int_{\mathbb{R}} g(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i), \qquad \sum_{x\in\mathbb{A}} g(x)P(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Monte-Carlo

Ley de los grandes números

Sean X_i variables aleatorias i.i.d. con esperanza finita $\mathbb{E}[X]$, entonces $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X]$ (w.p.1).

Método de Monte-Carlo

Sean X_i variables aleatorias i.i.d. con pdf $p(x)$ o pmf $P(x)$, entonces

$$\int_{\mathbb{R}} g(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i), \qquad \sum_{x\in\mathbb{A}} g(x)P(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Ideas Principales

En lugar de obtener la distribución a posteriori, vamos a generar muestras de esta distribución y aproximar la predictiva por Monte-Carlo.

$$p(x_{\text{test}}|\mathcal{D}_n) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(x_{\text{test}}|T_i)$$

Cadenas de Markov

Teorema de Ergódico

No es sencillo generar muestras independientes de la posterior sin una expresión analítica. El Teorema de Ergódico nos permite aplicar la ley de los grandes números, sin la hipótesis de independencia, para sucesiones de variables aleatorias con ciertas características.

$$\mathbb{E}[g(X)] \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Cadenas de Markov

Teorema de Ergódico

No es sencillo generar muestras independientes de la posterior sin una expresión analítica. El Teorema de Ergódico nos permite aplicar la ley de los grandes números, sin la hipótesis de independencia, para sucesiones de variables aleatorias con ciertas características.

$$\mathbb{E}[g(X)] \approx \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

Proceso de Markov

$$T_0 \to T_1 \to T_2 \to T_3 \to T_4 \to T_5 \to \cdots$$

Una **cadena de Markov** es una sucesión de variables aleatorias $\{T_t\}_{t \in \mathbb{N}_0}$ que cumple la propiedad de *falta de memoria*:

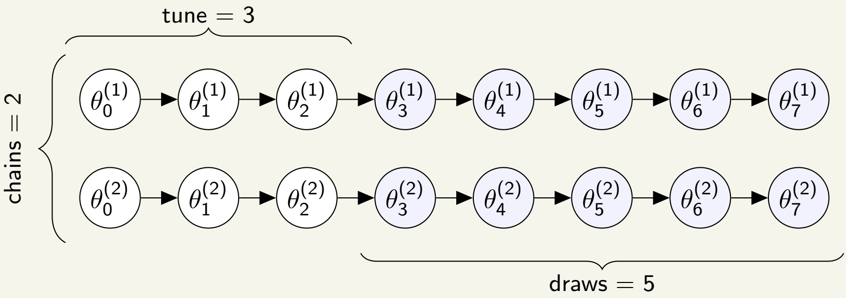
$$P(T_t = \theta \to T_{t+1} = \theta') = \pi(\theta' | \theta)$$

Monte Carlo Markov Chain (MCMC)

¿Que busco de una cadena de Markov?

- **Ergodicidad**: Que los promedios tiendan a las esperanzas
- **Estado Estacionario:** Que la distribución de las variables de la cadena alcancen el estado estacionario de la distribución deseada:

$$\pi(\theta') = \int_{\Theta} P(\theta \to \theta')\pi(\theta)d\theta$$



Técnicas de Muestreo

Muestreo de Gibbs

Muestreo utilizado cuando la conjunta $\pi(\alpha, \beta)$ es dificil de muestrear pero las condicionales, $\pi(\alpha|\beta)$ y $\pi(\beta|\alpha)$, son factibles. Itera entre:

$$\beta_k \sim \pi(\beta|\alpha_k), \qquad \alpha_{k+1} \sim \pi(\alpha|\beta_k)$$

Técnicas de Muestreo

Muestreo de Gibbs

Muestreo utilizado cuando la conjunta $\pi(\alpha, \beta)$ es dificil de muestrear pero las condicionales, $\pi(\alpha|\beta)$ y $\pi(\beta|\alpha)$, son factibles. Itera entre:

$$\beta_k \sim \pi(\beta|\alpha_k), \qquad \alpha_{k+1} \sim \pi(\alpha|\beta_k)$$

Muestreo Metrópolis

Muestreo utilizado cuando la dificultad de la posteriori radica en la constante de multiplicación $\pi(\theta) = \frac{f(\theta)}{Z}$. Itera entre

- Generar $\theta' = \theta_t + \delta$. Para el caso discreto, $\delta \sim \mathcal{U}\{-1, 0, 1\}$ (uniforme discreta de 3 átomos). Para el caso continuo $\delta \sim \mathcal{N}(0, \sigma^2)$.
- Sortear una Bernoulli de probabilidad $\alpha(\theta_t, \theta')$. Si dicha variable vale 1, $\theta_{t+1} = \theta'$; caso contrario $\theta_{t+1} = \theta_t$.

$$\alpha(\theta_a, \theta_b) = \min\left\{1, \frac{f(\theta_b)}{f(\theta_a)}\right\}$$

Técnicas de Muestreo

Algorithm 6 No-U-Turn Sampler with Dual Averaging

Given θ^0, δ, $\mathcal{L}, M, M^{\text{adapt}}$:
Set $\epsilon_0 = \text{FindReasonableEpsilon}(\theta), \mu = \log(10\epsilon_0), \bar{\epsilon}_0 = 1, \bar{H}_0 = 0, \gamma = 0.05, t_0 = 10, \kappa = 0.75.$
for $m = 1$ to M **do**
  Sample $r^0 \sim \mathcal{N}(0, I)$.
  Resample $u \sim \text{Uniform}([0, \exp\{\mathcal{L}(\theta^{m-1} - \frac{1}{2}r^0 \cdot r^0\}])$
  Initialize $\theta^- = \theta^{m-1}$, $\theta^+ = \theta^{m-1}$, $r^- = r^0, r^+ = r^0, j = 0, \theta^m = \theta^{m-1}, n = 1, s = 1.$
  while $s = 1$ **do**
    Choose a direction $v_j \sim \text{Uniform}(\{-1, 1\})$.
    if $v_j = -1$ **then**
      $\theta^-, r^-, -, -, \theta', n', s', \alpha, n_\alpha \leftarrow \text{BuildTree}(\theta^-, r^-, u, v_j, j, \epsilon_{m-1}\theta^{m-1}, r^0)$.
    else
      $-, -, \theta^+, r^+, \theta', n', s', \alpha, n_\alpha \leftarrow \text{BuildTree}(\theta^+, r^+, u, v_j, j, \epsilon_{m-1}, \theta^{m-1}, r^0)$.
    end if
    if $s' = 1$ **then**
      With probability $\min\{1, \frac{n'}{n}\}$, set $\theta^m \leftarrow \theta'$.
    end if
    $n \leftarrow n + n'$.
    $s \leftarrow s'\mathbb{I}[(\theta^+ - \theta^-) \cdot r^- \geq 0]\mathbb{I}[(\theta^+ - \theta^-) \cdot r^+ \geq 0]$.
    $j \leftarrow j + 1$.
  end while
  if $m \leq M^{\text{adapt}}$ **then**
    Set $\bar{H}_m = \left(1 - \frac{1}{m+t_0}\right) \bar{H}_{m-1} + \frac{1}{m+t_0}(\delta - \frac{\alpha}{n_\alpha})$.
    Set $\log \epsilon_m = \mu - \frac{\sqrt{m}}{\gamma}\bar{H}_m, \log \bar{\epsilon}_m = m^{-\kappa} \log \epsilon_m + (1 - m^{-\kappa}) \log \bar{\epsilon}_{m-1}$.
  else
    Set $\epsilon_m = \bar{\epsilon}_{M^{\text{adapt}}}$.
  end if
end for

No-U-Turn Sampler (NUTS)

Muestreo para variables continuas con posterior diferenciable. Explora soluciones sobre curvas de nivel de

$$-\log \pi(\theta) + \|r\|^2$$

Hoffman - Gelman 2014: "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo".

Calidad de las muestras

Tamaño efectivo de la muestra: ESS bulk

El teorema ergódico permite aproximar esperanzas a partir de promedios sin la hipótesis de independencia, pero con una convergencia más lenta. El tamaño efectivo de la muestra (Effective Sample Size, ESS) es la cantidad de datos independientes necesarios para alcanzar la misma varianza que posee el promedio de las muestras.

Calidad de las muestras

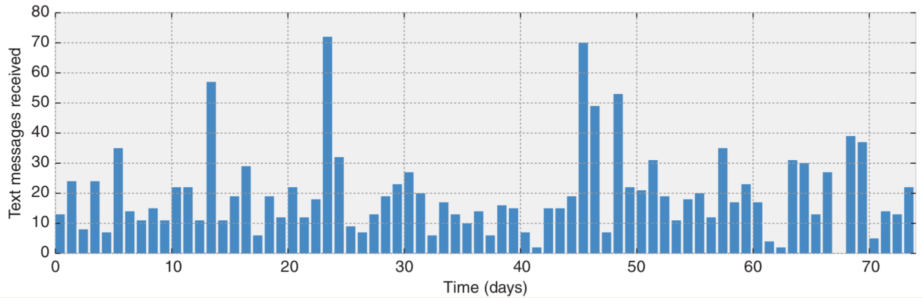
Tamaño efectivo de la muestra: ESS bulk

El teorema ergódico permite aproximar esperanzas a partir de promedios sin la hipótesis de independencia, pero con una convergencia más lenta. El tamaño efectivo de la muestra (Effective Sample Size, ESS) es la cantidad de datos independientes necesarios para alcanzar la misma varianza que posee el promedio de las muestras.

Diagnóstico Gelman-Rubin: $\hat{R}$

Tener muchas cadenas de experimento nos permite corroborar si se alcanzó el estado estacionario. Si todas convergieron a la misma distribución, entonces la varianza entre cadenas debería ser similar a la varianza dentro de cada cadena. Se denomina $\hat{R}$ al cociente entre estas varianzas. Suele considerarse $\hat{R} > 1.01$ una señal de alerta y un valor $\hat{R} > 1.1$ un problema a resolver.

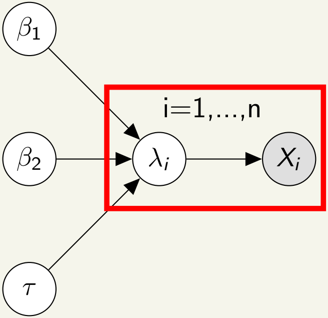
Ejemplo de modelado Bayesiano



Problema

Un usuario proporciona una serie de recuentos diarios de mensajes de whatsapp enviados. Tiene curiosidad por saber si los hábitos de envío de mensajes han cambiado con el tiempo. ¿Cómo puedes modelar esto?

PYMC: Programación Probabilística con MCMC



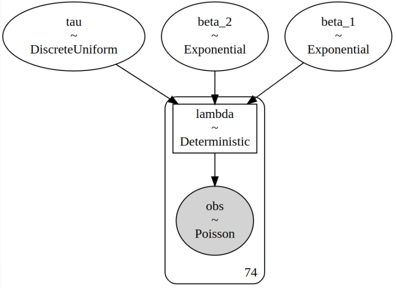
- $\beta_1 \sim \mathcal{E}(\alpha)$
- $\beta_2 \sim \mathcal{E}(\alpha)$
- $\tau \sim \mathcal{U}\{1 : T\}$
- $\lambda_i = \begin{cases} \beta_1 & i < \tau \\ \beta_2 & i \geq \tau \end{cases}$
- $X_i|_{\lambda_i} \sim \text{Poi}(\lambda_i)$

PYMC

```python
import pymc as pm
import numpy as np
import matplotlib.pyplot as plt

count_data = np.loadtxt("txtdata.csv")
n_count_data = len(count_data)
with pm.Model() as model:
    alpha = 1.0/count_data.mean()
    beta_1 = pm.Exponential("beta_1", alpha)
    beta_2 = pm.Exponential("beta_2", alpha)
    tau = pm.DiscreteUniform("tau", lower=0, upper=n_count_data - 1)
    idx = np.arange(n_count_data) # Index
    lambda_ = pm.Deterministic("lambda",pm.math.switch(tau > idx, beta_1, beta_2)
    observation = pm.Poisson("obs", lambda_, observed=count_data)
    trace = pm.sample(draws=1000, tune=1000, chains=2)

pm.model_to_graphviz(model)
```

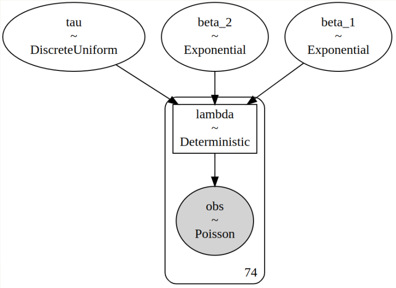


PYMC

```python
import pymc as pm
import numpy as np
import matplotlib.pyplot as plt

count_data = np.loadtxt("txtdata.csv")
n_count_data = len(count_data)
with pm.Model() as model:
    alpha = 1.0/count_data.mean()
    beta_1 = pm.Exponential("beta_1", alpha)
    beta_2 = pm.Exponential("beta_2", alpha)
    tau = pm.DiscreteUniform("tau", lower=0, upper=n_count_data - 1)
    idx = np.arange(n_count_data) # Index
    lambda_ = pm.Deterministic("lambda",pm.math.switch(tau > idx, beta_1, beta_2))
    observation = pm.Poisson("obs", lambda_, observed=count_data)
    trace = pm.sample(draws=1000, tune=1000, chains=2)

pm.model_to_graphviz(model)
```




```python
import arviz as az

summary = az.summary(trace, var_names=["beta_1","beta_2","tau"])
print(summary)
```

```
          mean     sd  ess_bulk  r_hat
beta_1  17.772  0.630    1834.0    1.0
beta_2  22.677  0.887    1515.0    1.0
tau     44.287  0.863     247.0    1.0
```

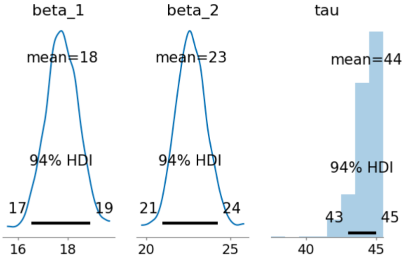
PYMC

```python
beta_1_samples = trace.posterior['beta_1'].values
beta_2_samples = trace.posterior['beta_2'].values
tau_samples = trace.posterior['tau'].values
lambda_samples = trace.posterior['lambda'].values

_ = pm.plot_posterior(trace.posterior[['beta_1','beta_2','tau']],figsize=(7,4))

with model:
  posterior_pred = pm.sample_posterior_predictive(trace,predictions=True)

pred_samples = posterior_pred.predictions['obs'].values
```



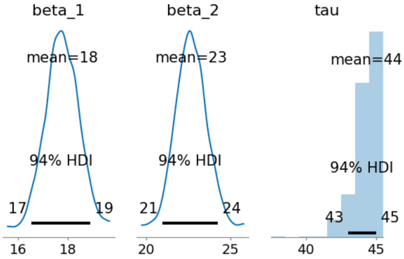
PYMC

```
beta_1_samples = trace.posterior['beta_1'].values
beta_2_samples = trace.posterior['beta_2'].values
tau_samples = trace.posterior['tau'].values
lambda_samples = trace.posterior['lambda'].values

_ = pm.plot_posterior(trace.posterior[['beta_1','beta_2','tau']],figsize=(7,4))

with model:
  posterior_pred = pm.sample_posterior_predictive(trace,predictions=True)

pred_samples = posterior_pred.predictions['obs'].values
```




```
plt.bar(np.arange(n_count_data), pred_samples[0,-1], color="#348ABD")
plt.xlabel("Tiempo (días)")
plt.ylabel("Cantidad de mensajes")
plt.xlim(-1, n_count_data)
plt.ylim(0, 75)
plt.show()
```